Knowledge Discovery & Data Mining

— Feature Analysis —

Non-Linear Relationships

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

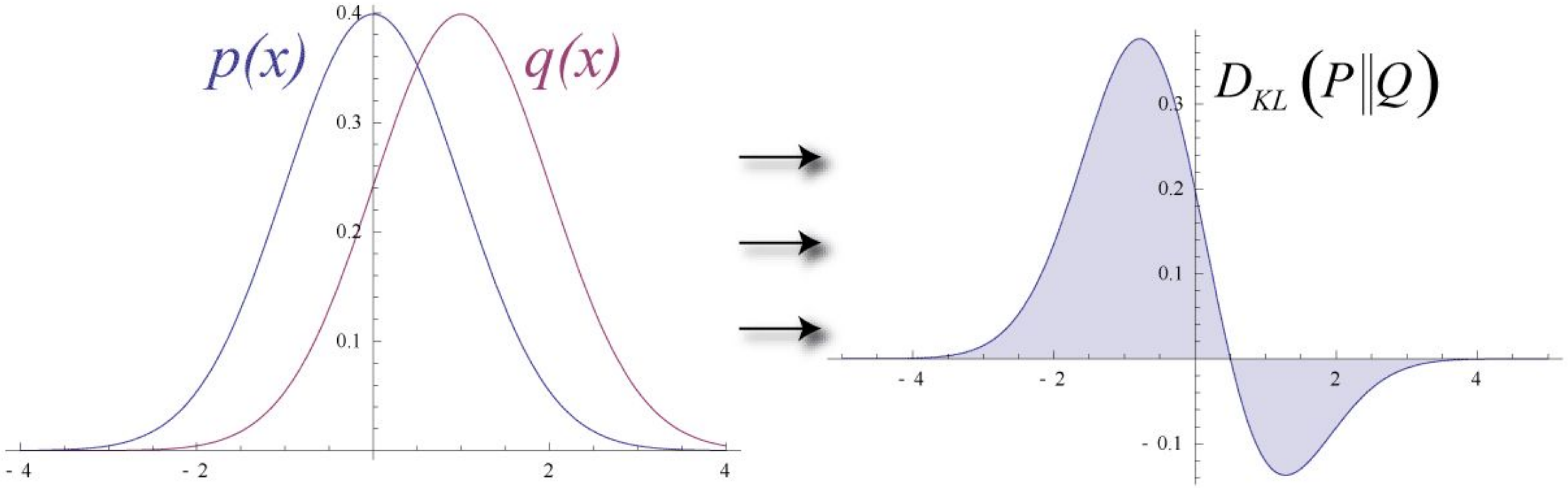
Recall: Analyzing Feature Relationships

- Analyzing Feature Relationships
  - Introduction to Feature Analysis
  - Covariance (for numerical features)
  - Correlation Coefficient (for numerical features)
  - Spearman's Rank Correlation (Numeric & Ordinal Data)
  - Chi-Square Test (for categorical features)
  - Partial correlation

Kullback–Leibler (KL) divergence

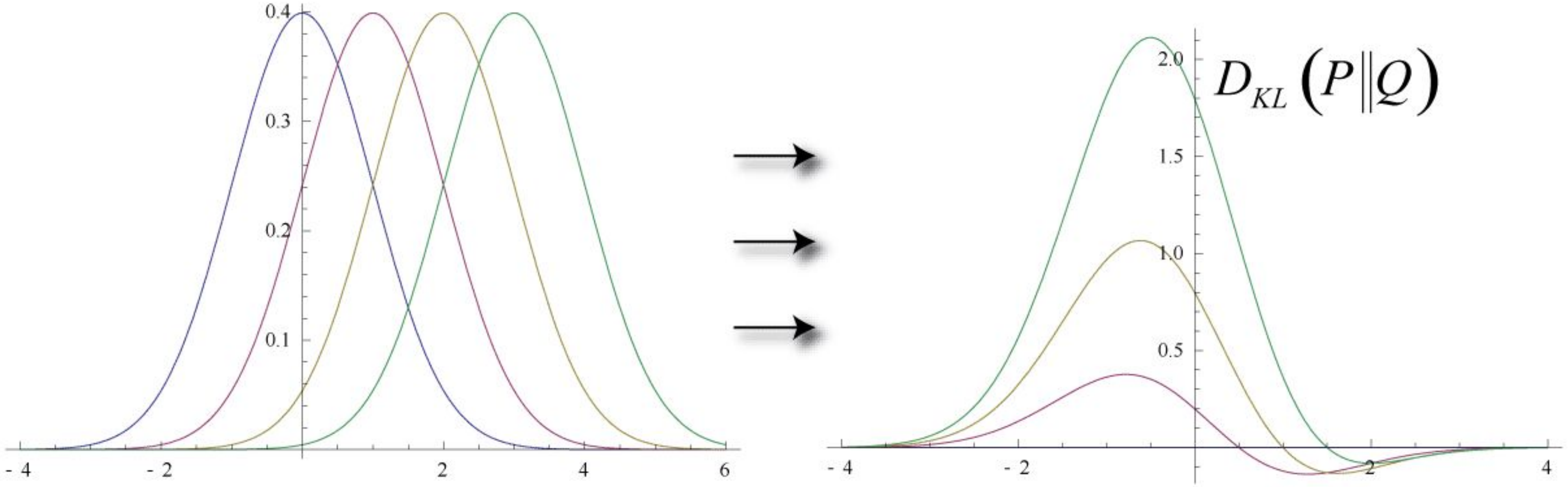
Also called relative entropy.

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$



Original Gaussian PDF's

KL Area to be Integrated



Mutual Information

Formally, the mutual information[1] of two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$

where $p(x, y)$ is the joint probability function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively.

Mutual Information

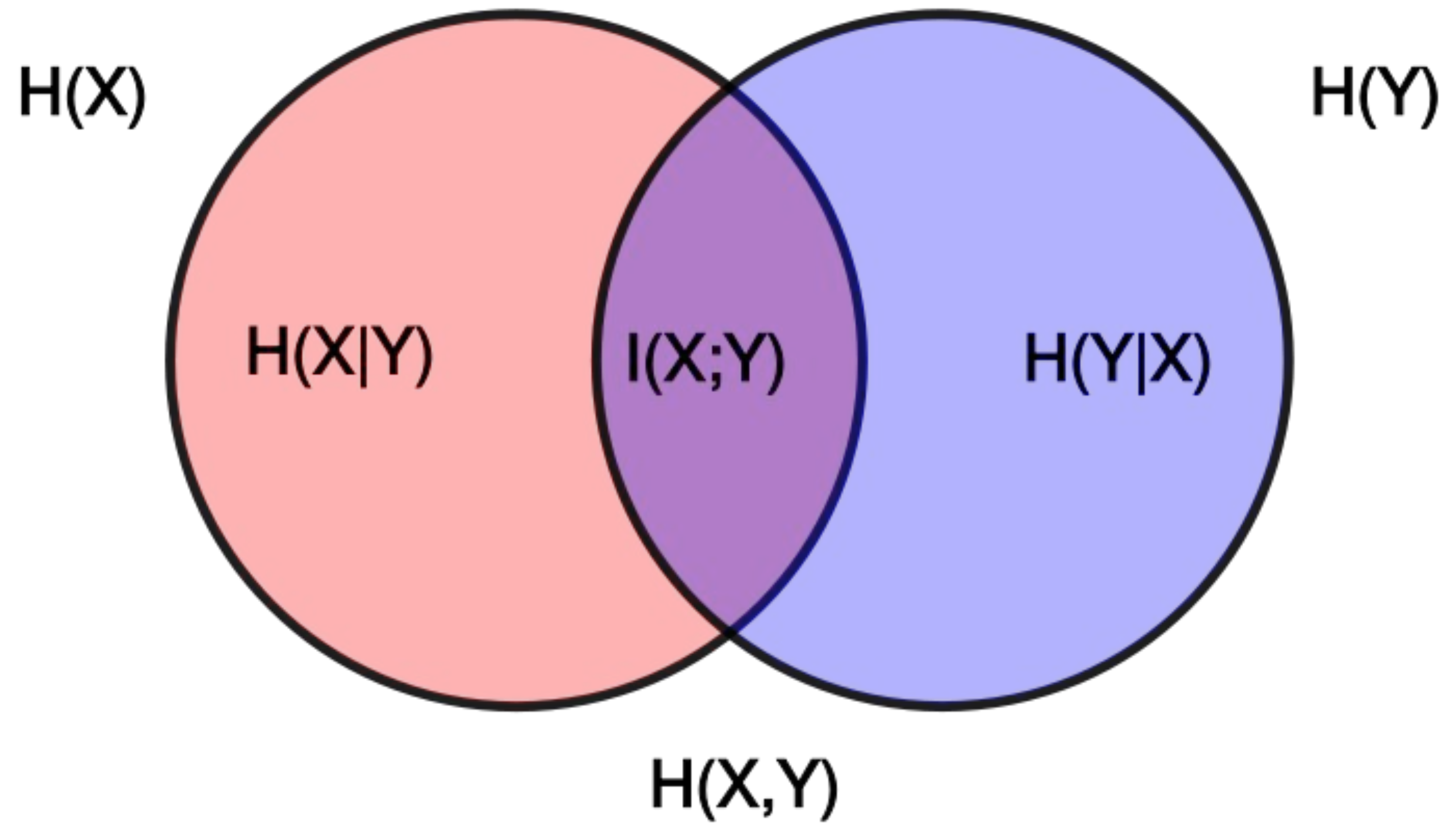
Mutual information can be equivalently expressed as

$$\begin{aligned} I(X;Y) &\equiv \mathrm{H}(X) - \mathrm{H}(X|Y) \\ &\equiv \mathrm{H}(Y) - \mathrm{H}(Y|X) \\ &\equiv \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \\ &\equiv \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \end{aligned}$$

where $\mathbf{H}(X)$ and $\mathbf{H}(Y)$ are the marginal entropies, $\mathrm{H}(X|Y)$ and $\mathrm{H}(Y|X)$ are the conditional entropies, and $\mathrm{H}(X,Y)$ is the joint entropy of X and Y.

Mutual Information

Diagram showing additive and subtractive relationships for various information measures associated with correlated variables X and Y. The area contained by both circles is the joint entropy H(X,Y). The circle on the left (red and violet) is the individual entropy H(X), with the red being the conditional entropy H(X|Y). The circle on the right (blue and violet) is H(Y), with the blue being H(Y|X). The violet is the mutual information I(X;Y).



Mutual Information

Mutual information can also be expressed as a Kullback–Leibler divergence of the product of the marginal distributions, $p(x) \times p(y)$, of the two random variables X and Y, from the random variables's joint distribution, $p(x, y)$:

$$I(X;Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y)).$$

Conditional Mutual Information

Definition The *conditional mutual information* of random variables X and Y given Z is defined by

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{2.60}$$

$$= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}. \tag{2.61}$$

Pairwise Nonlinear correlation

In statistics, the maximal information coefficient (MIC) is a measure of the strength of the linear or nonlinear association between two variables X and Y.

--- Wikipedia entry

Maximal Information Coefficient (MIC)

Definition Let D be a set of ordered pairs. For a grid G, let $D|_G$ denote the probability distribution induced by the data D on the cells of G, and let $I(-)$ denote mutual information. Let $I^*(D, x, y) = \max_G I(D|_G)$, where the maximum is taken over all x-by-y grids G (possibly with empty rows/columns). MIC is defined as

$$\text{MIC}(D) = \max_{xy<B(|D|)} \frac{I^*(D, x, y)}{\log_2 \min\{x, y\}}$$

Where B is a growing function satisfying $B(n) = o(n)$.

MIC

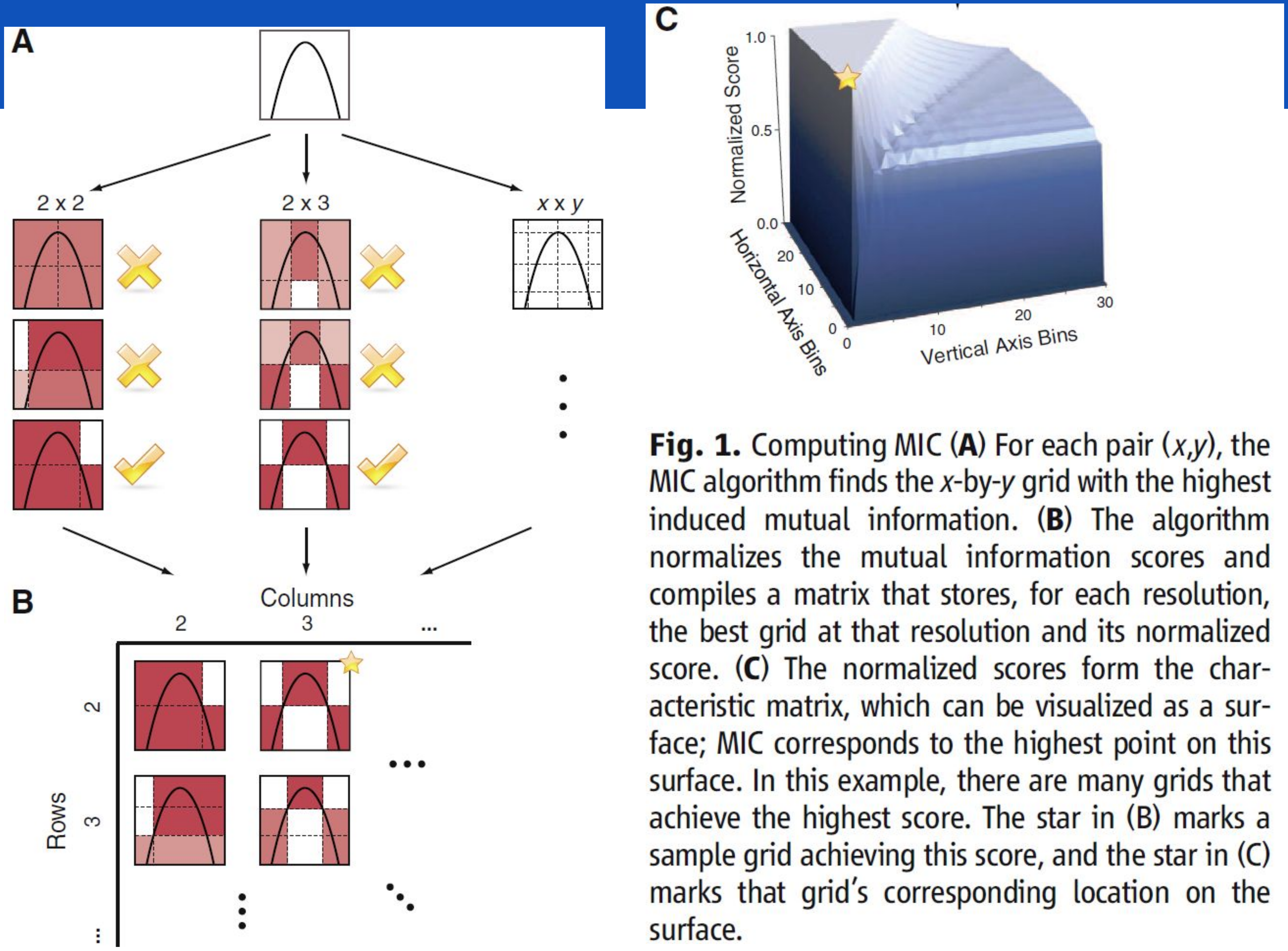


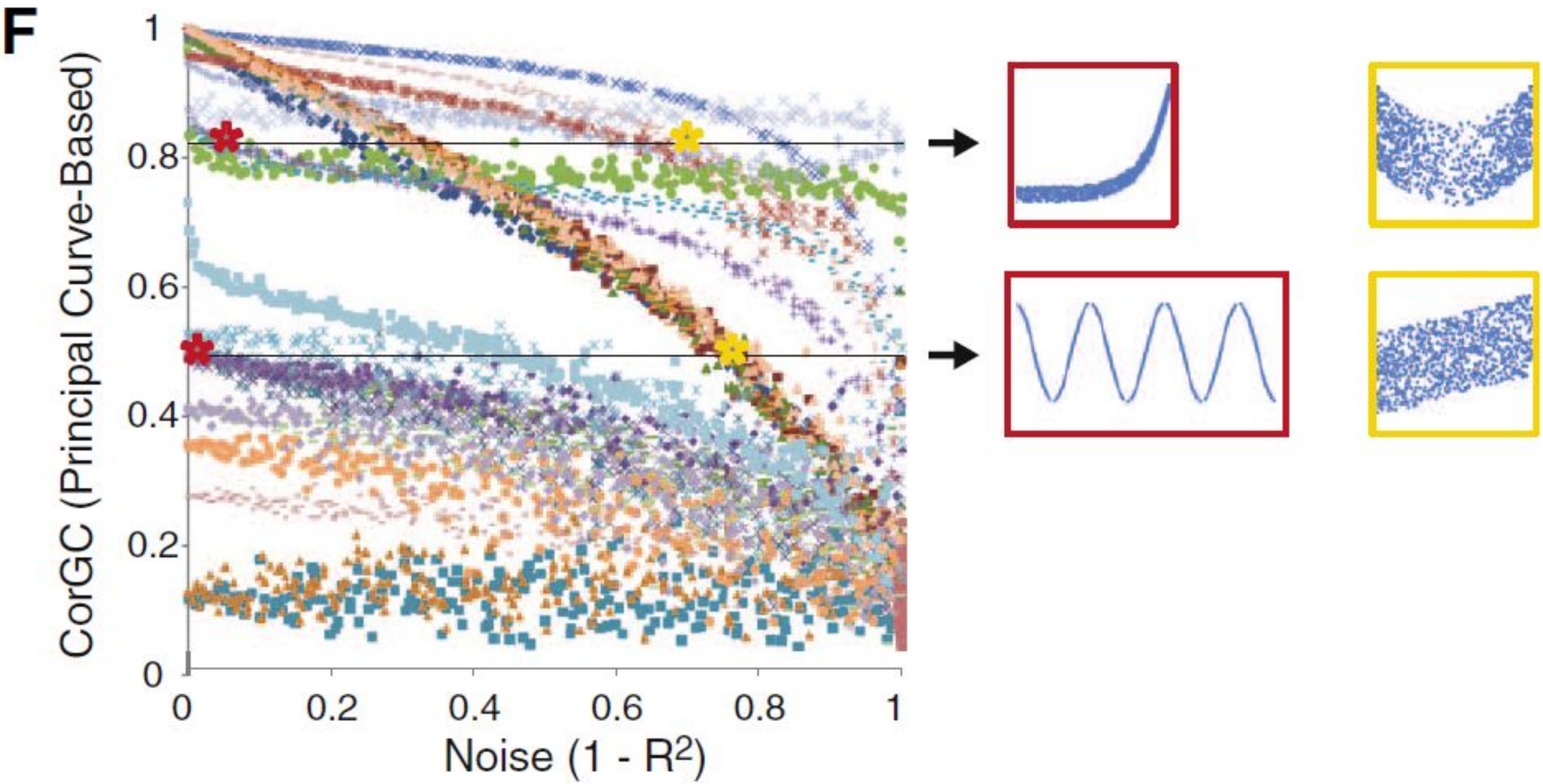
Fig. 1. Computing MIC (**A**) For each pair (*x*,*y*), the MIC algorithm finds the *x*-by-*y* grid with the highest induced mutual information. (**B**) The algorithm normalizes the mutual information scores and compiles a matrix that stores, for each resolution, the best grid at that resolution and its normalized score. (**C**) The normalized scores form the characteristic matrix, which can be visualized as a surface; MIC corresponds to the highest point on this surface. In this example, there are many grids that achieve the highest score. The star in (B) marks a sample grid achieving this score, and the star in (C) marks that grid's corresponding location on the surface.

MIC

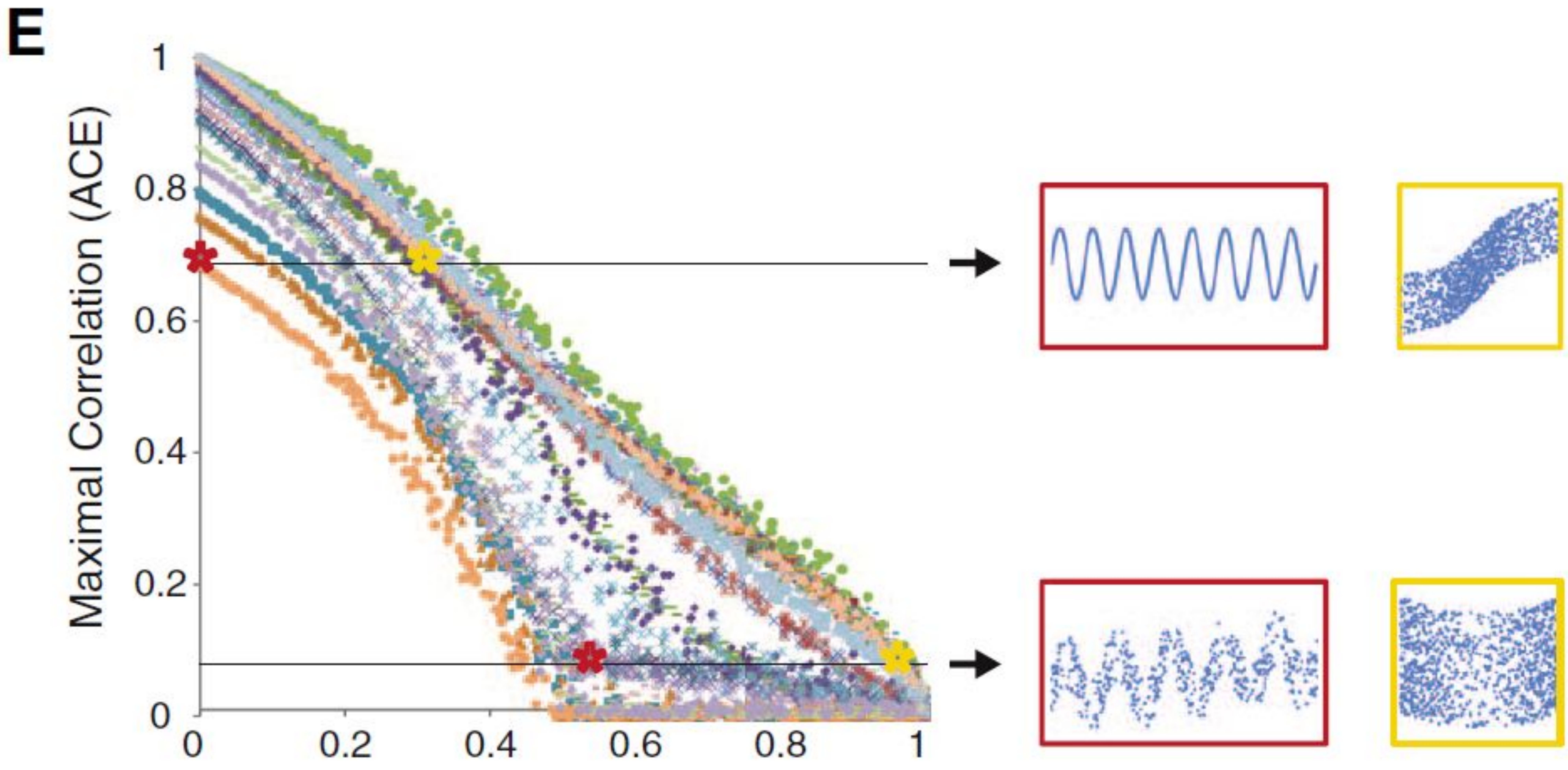
A

Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE)	Mutual Information (Kraskov)	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal *(Fourier frequency)*	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal *(non-Fourier frequency)*	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal *(varying frequency)*	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

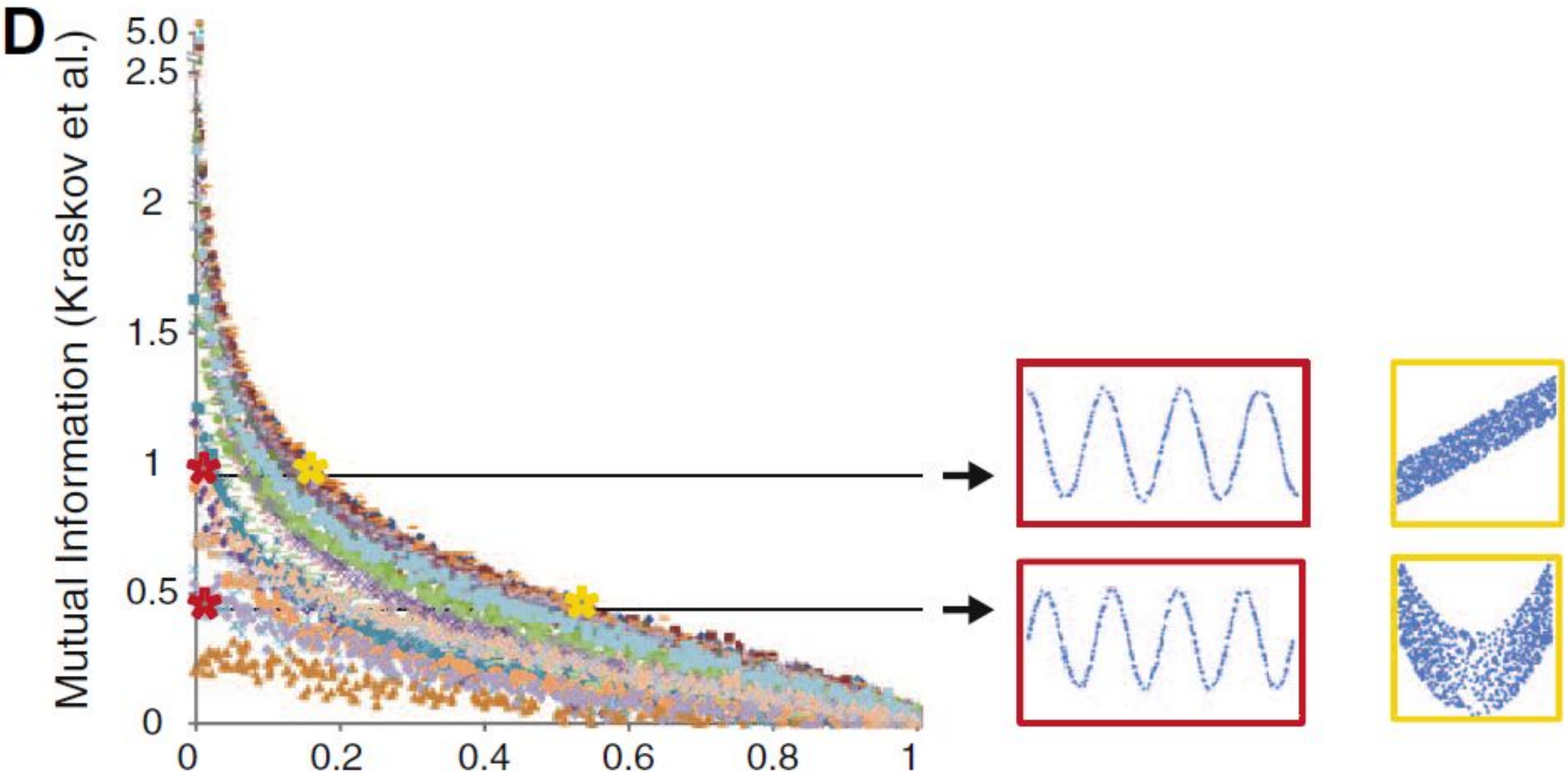
MIC



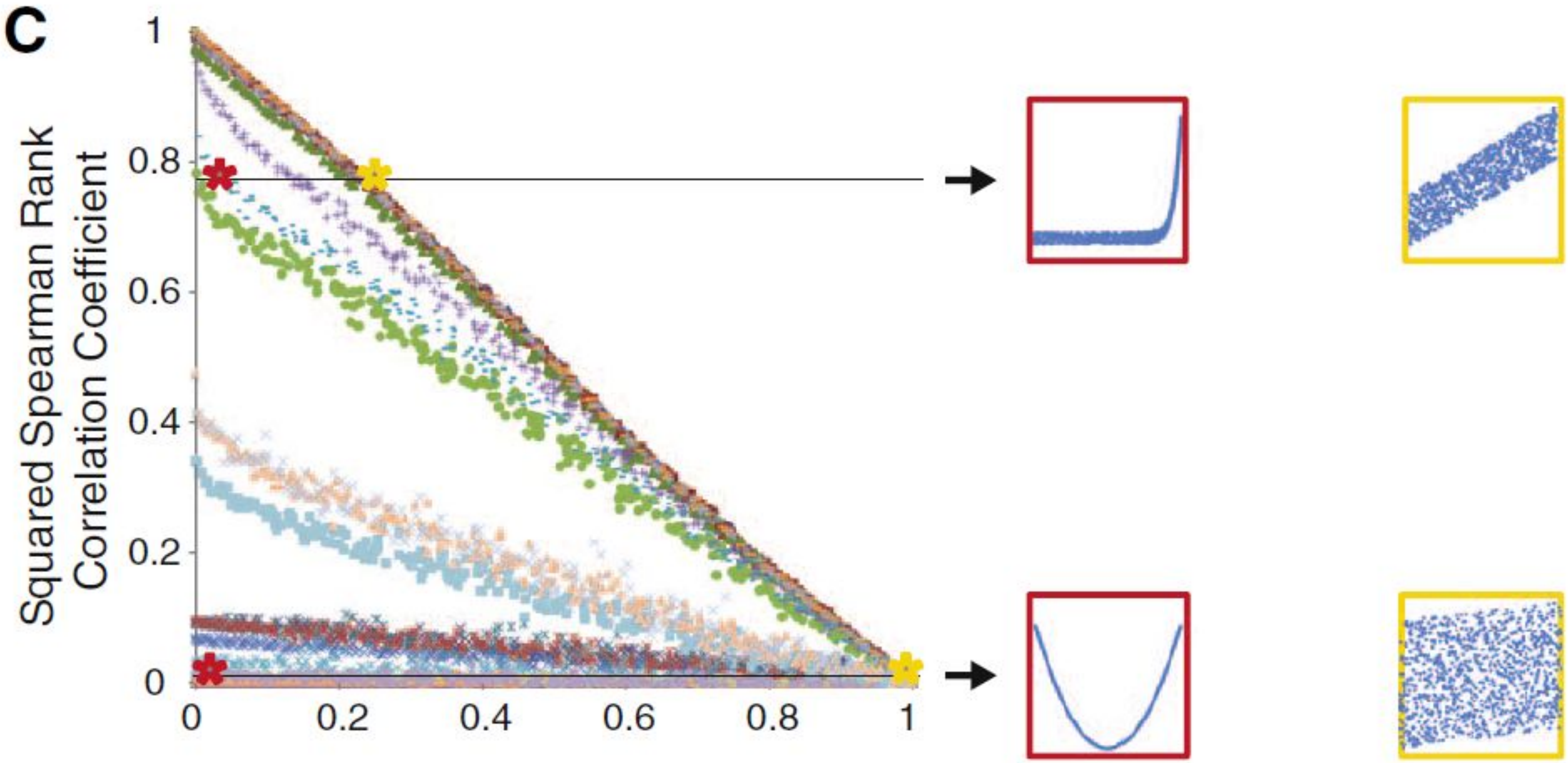
MIC



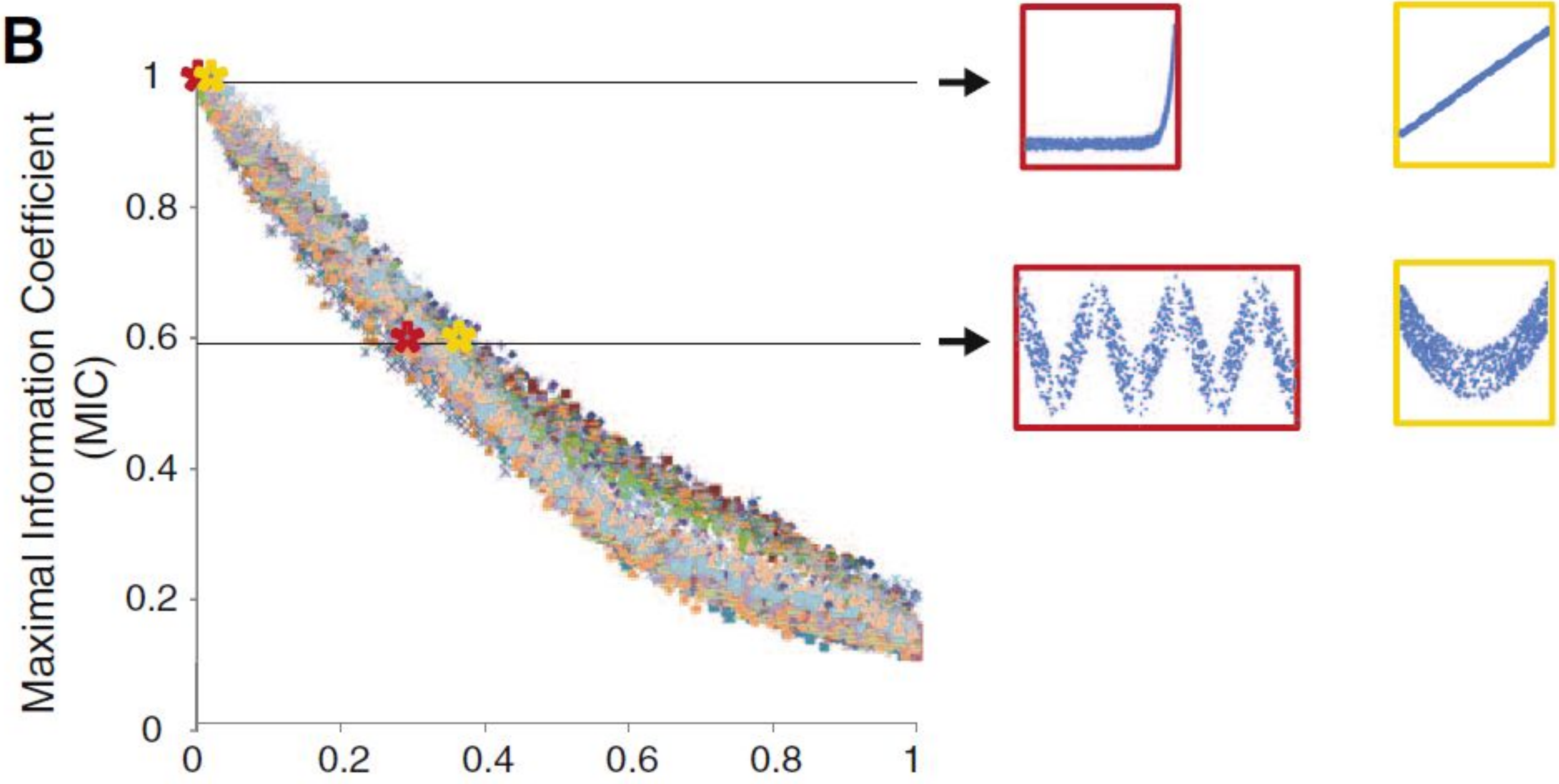
MIC


D
Mutual Information (Kraskov et al.)
5.0
2.5
2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1

MIC

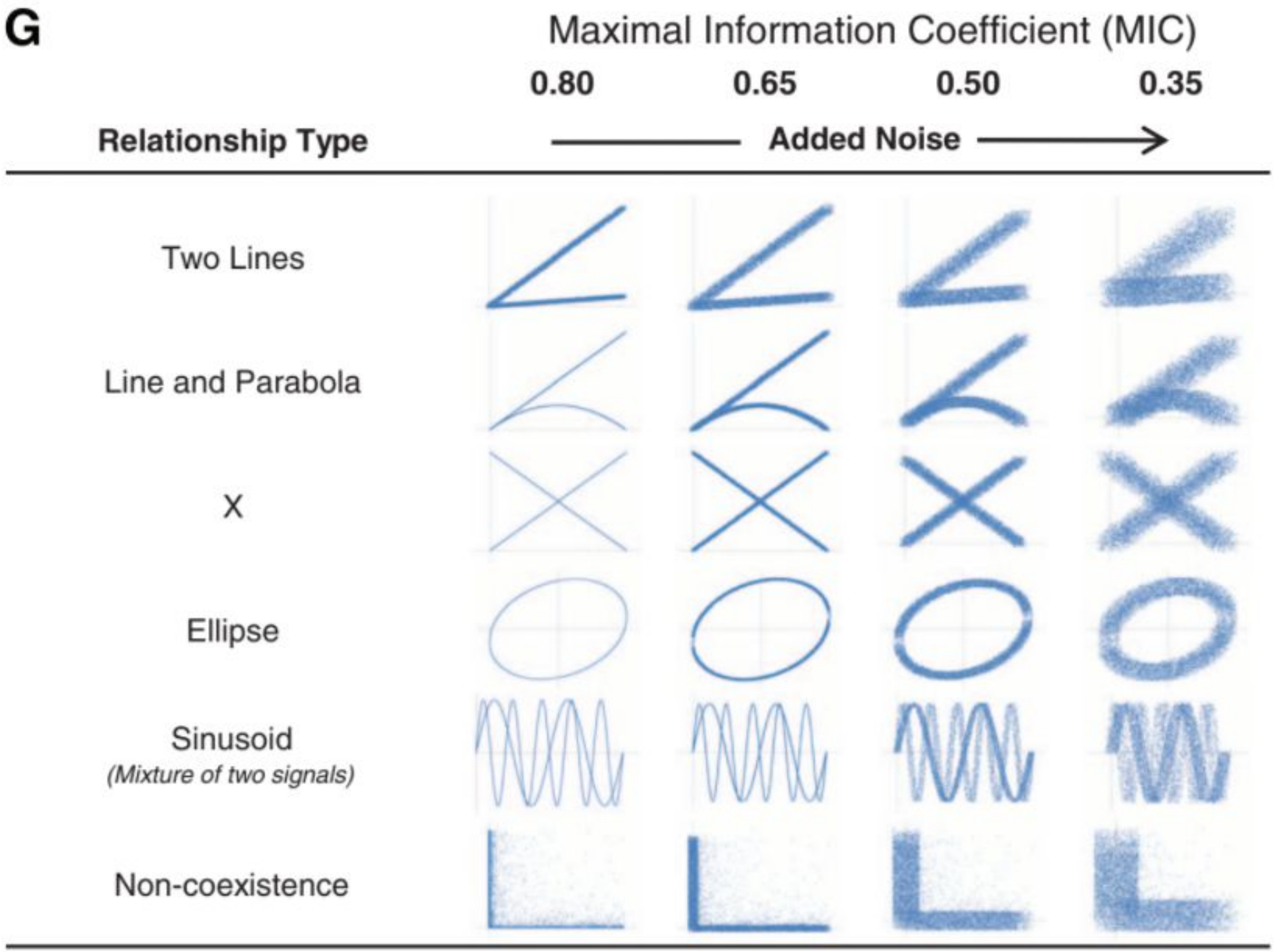

C
Squared Spearman Rank Correlation Coefficient
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

MIC



B
Maximal Information Coefficient
(MIC)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

MIC


G
Maximal Information Coefficient (MIC)
0.80
0.65
0.50
0.35
Relationship Type
Added Noise
Two Lines
Line and Parabola
X
Ellipse
Sinusoid
(Mixture of two signals)
Non-coexistence

MIC

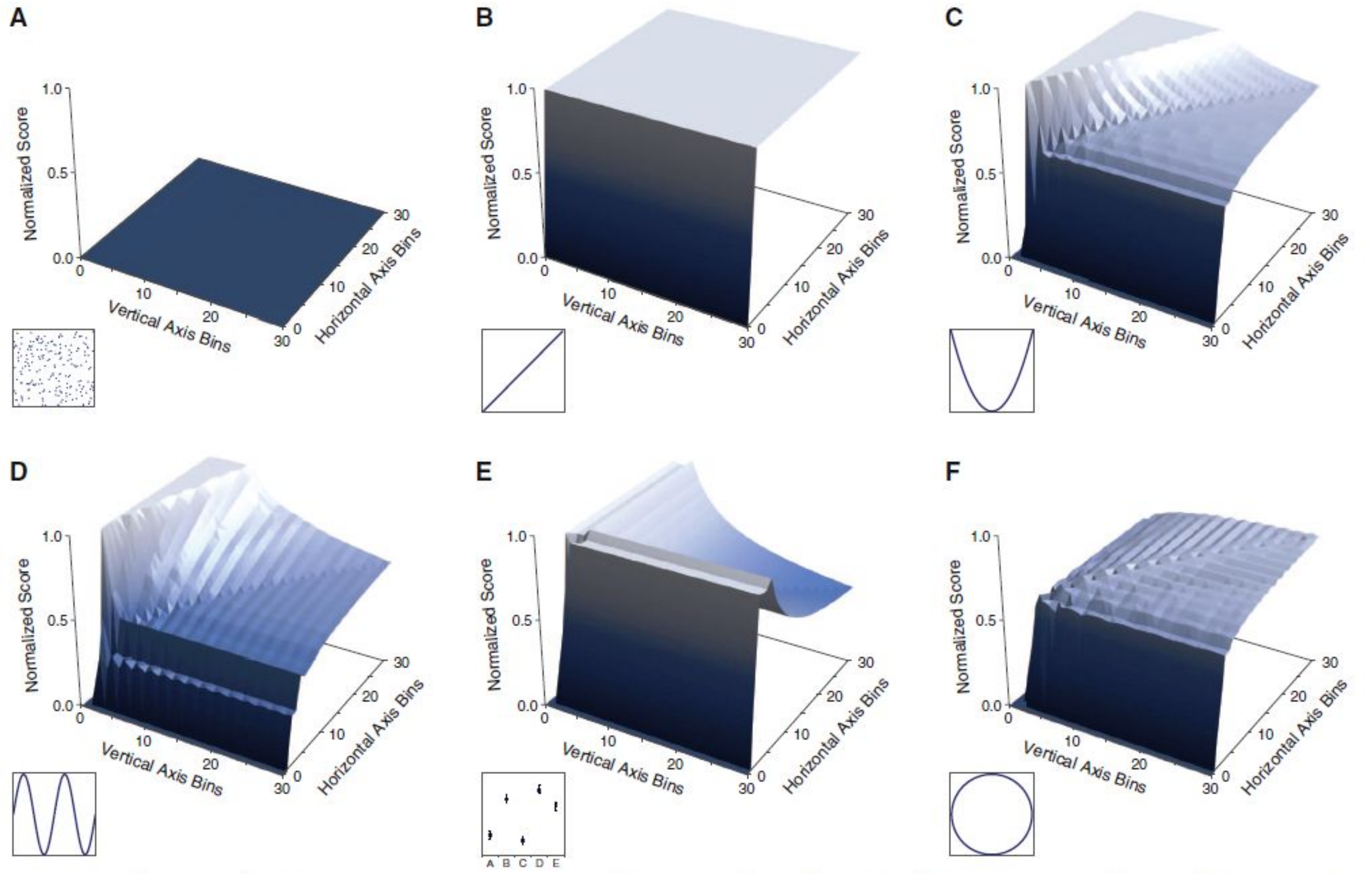
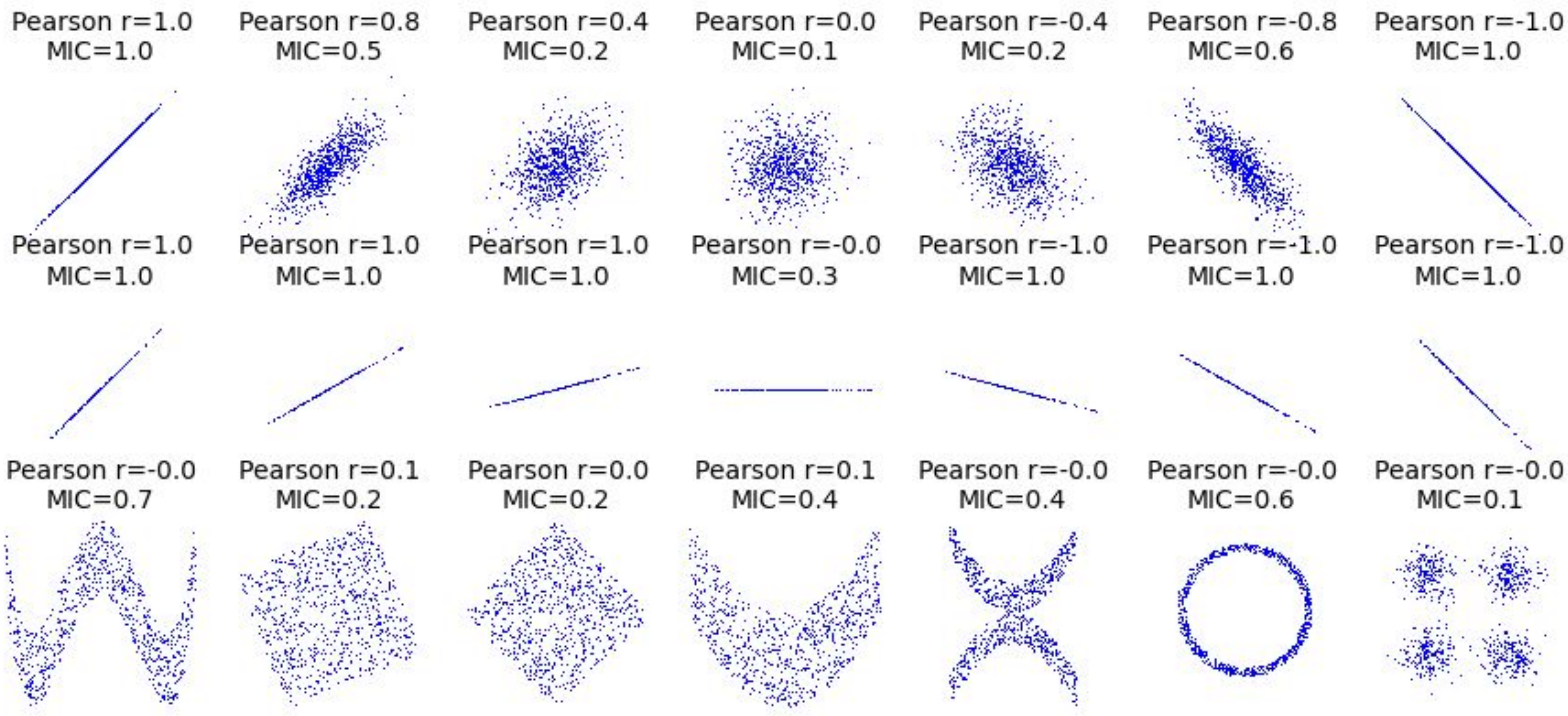


Fig. 3. Visualizations of the characteristic matrices of common relationships. (**A** to **F**) Surfaces representing the characteristic matrices of several common relationship types. For each surface, the *x* axis represents number of vertical axis bins (rows), the *y* axis represents number of horizontal axis bins (columns), and the *z* axis represents the normalized score of the best-performing grid with those dimensions. The inset plots show the relationships used to generate each surface. For surfaces of additional relationships, see fig. S7.

MIC


Pearson r=1.0
MIC=1.0
Pearson r=0.8
MIC=0.5
Pearson r=0.4
MIC=0.2
Pearson r=0.0
MIC=0.1
Pearson r=-0.4
MIC=0.2
Pearson r=-0.8
MIC=0.6
Pearson r=-1.0
MIC=1.0
Pearson r=1.0
MIC=1.0
Pearson r=1.0
MIC=1.0
Pearson r=1.0
MIC=1.0
Pearson r=-0.0
MIC=0.3
Pearson r=-1.0
MIC=1.0
Pearson r=-1.0
MIC=1.0
Pearson r=-1.0
MIC=1.0
Pearson r=-0.0
MIC=0.7
Pearson r=0.1
MIC=0.2
Pearson r=0.0
MIC=0.2
Pearson r=0.1
MIC=0.4
Pearson r=-0.0
MIC=0.4
Pearson r=-0.0
MIC=0.6
Pearson r=-0.0
MIC=0.1

Limitation of MIC

Consider a toy data set with three dimensions {A, B, C}. MIC can find two separate ways to discretize B to maximize its correlation with A and C, but it cannot find a discretization of B such that the correlation with regard to both A and C is maximized. Thus, MIC is not suited for calculating correlations over more than two dimensions. Further, adapting existing solutions to the multivariate setting is nontrivial due to the huge search space.

Reference

- Detecting Novel Associations in Large Data Sets
- Multivariate Maximal Correlation Analysis